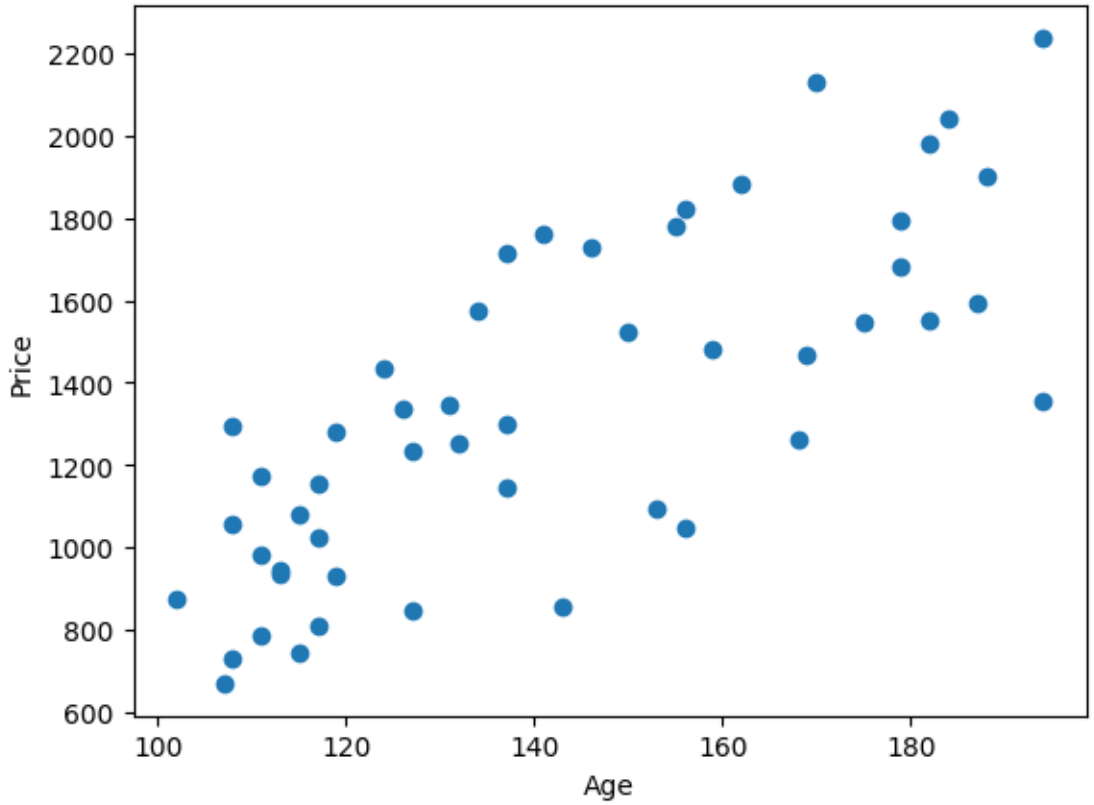
The text file Wine.csv contains the Price (in dollars) of 50 wine bottles in different distilleries along with the Age of the wine in years and the Alcohol %. Answer the following:

1. **How is Age of Wine and Alcohol % affecting Price?**

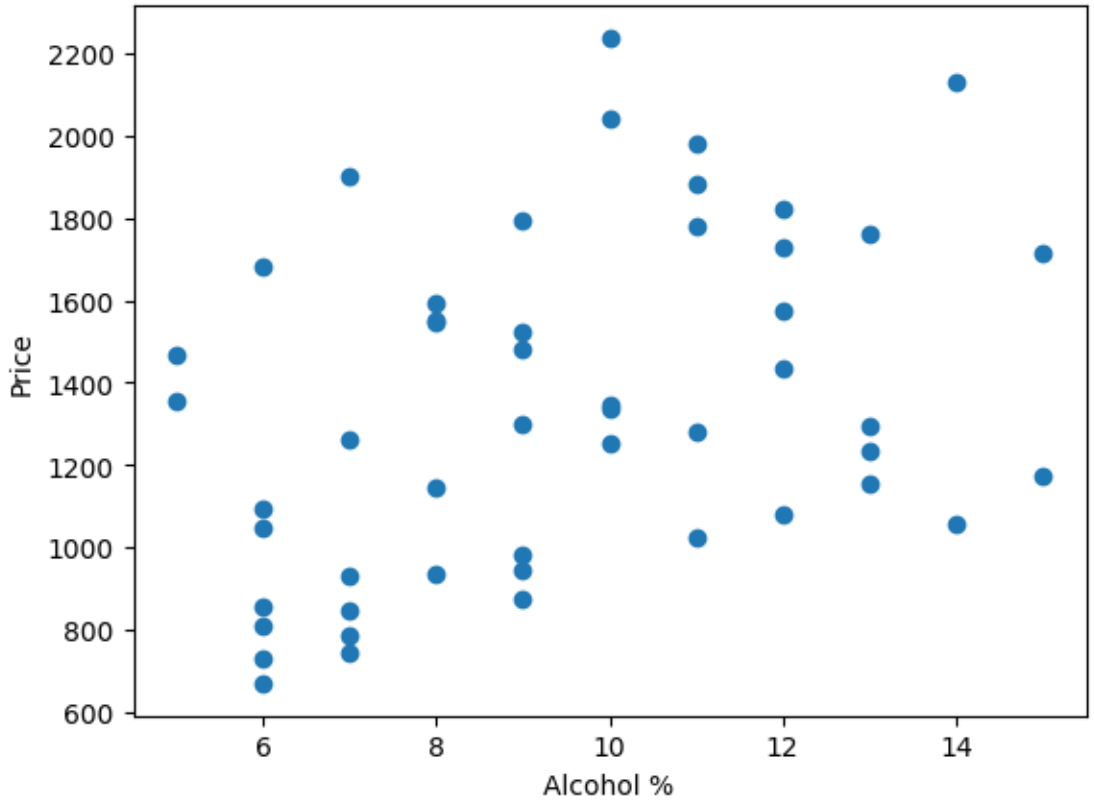
    - To analyse how the Age of Wine and Alcohol % affect Price, we can create a scatter plot for each variable, where the x-axis represents Age or Alcohol % and the y-axis represents Price.
    - We can also calculate the correlation coefficient between each variable and Price to determine the strength and direction of their relationship.

Age vs Price



From the plot, we can observe that there is a **positive linear relationship** between the Age (Independent Variable) and Price (Dependent Variable) of column.

Alcohol Percent vs Price



From the plot, we can observe that there is **no strong relationship** between the Alcohol (Independent Variable) and Price (Dependent Variable) of column

Correlation Coefficient:

```
# calculate the correlation coefficient between Age and Price
age_corr = data['Age'].corr(data['Price'])
print(f"Correlation coefficient between Age and Price: {age_corr:.2f}")
```

```
Correlation coefficient between Age and Price: 0.76
```

```
# calculate the correlation coefficient between Alcohol % and Price
alcohol_corr = data['Alcohol_percent'].corr(data['Price'])
print(f"Correlation coefficient between alcohol% and Price: {alcohol_corr:.2f}")
```

```
Correlation coefficient between alcohol% and Price: 0.38
```

Already there is positive strong relationship between Age and Price (from scatter plot) and **Correlation coefficient also 0.76** to so we can say there is positive relationship between them.

2. **Using a first order multiple regression model to the data and answer the following:**

A. Is the model useful?

- The dataset contains only 50 records which reduces the chance of overfitting model
- The R2 (R- squared) score is 0.82. This indicates that the model fits well on data and may give accurate predictions. (Above 0.9 considered to be good model)
- The other Evaluation metrics:
  Mean Absolute Error: 118.43,
  Mean Squared Error: 18391.97,
  Root Mean Squared Error: 135.51
  In General, Lower the value of metrics indicates better the model performance

B. Given the age of wine, by what amount can one expect the price to go up for an increase in Alcohol of 1%?

- These coefficients are estimates obtained from performing multiple regression model
  Intercept: -1330.86
  Age coefficient: 12.78
  Alcohol Percent coefficient: 89.90

To find the price of wine when there is an increase in 1% (keeping age coefficient constant)

$Y = ß_o + ß_1x_1 + ß_2x_2$

$Y = 0 + 1(89.90) + \text{constant}$

$Y = 93$

However, **if the Alcohol Percent of a wine increases by one percentage point, the estimated price of the wine would increase by $93, holding the Age constant.**

C. You find a wine bottle in the distillery that is 120 years old and find that it contains 15% Alcohol. What is the minimum amount you should be willing to pay so that you are 97% percent certain to buy the bottle?

```
import numpy as np

# Predict the price of a bottle with age 120 years and alcohol percentage 15%
x_new = np.array([[120, 15]])
pred = regressor.predict(x_new)
pred
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have
  warnings.warn(
array([1552.15895435])
```

The minimum amount willing to pay for buy a bottle is $1552

E. In presence of the other, which of the two factors, Age of the bottle or Alcohol content, is more important in determining the selling price of a wine bottle?

- To determine which factor, Age of the bottle or Alcohol content, is more important in determining the selling price of a wine bottle, we can compare the magnitudes of their coefficients obtained from a linear regression model that includes both predictors.
- Alcohol Percent coefficient (92.99) is higher than Age coefficient (12.5)
- So, **Alcohol Percent is more important in determining the selling price of a wine bottle**, since the magnitude of its coefficient is larger than that of the Age coefficient.

3. Is there merit in trying higher order multiple regression models? Does the model fit improve?

- Trying higher order multiple regression models, **increases the complexity and may increase the risk of overfitting** (while increasing the degrees)
- Sometimes higher order multiple regression model show improvement in fitting data than the simpler models. However, it is important to note that the more complex model will difficult to interpret
- Before trying higher order models, it is important to check for any violations of the assumptions of regression model, such as linearity, normality, homoscedasticity, and independence of errors. If there any violations these should be addressed before attempting complex models